

To begin, we aimed to use different models on a few datasets in order to see which datasets were best fit by certain models. Of course, this meant we must have a way of comparing them. The four metrics we used to assess our models were r^2 , adjusted r^2 , r^2 cross validated, and AIC. r^2 is a metric that tells us how much of the variability in the data can be explained by our model, and works on a scale of 0 to 1. Because of the nature of r^2 (being $1 - \frac{\text{sum of squares from the regression}}{\text{total sum of squares}}$), it can never decrease as you add more predictors. Because of that, we need adjusted r^2 and r^2 cross validated, two metrics that are similar to r^2 , but add some penalization for adding poor predictors. Lastly, we have AIC, a metric that is more known for its penalization than for rewarding for good predictors. Overall this gives us a nice balance of metrics to use, that while rewarding us for adding quality predictors to the model, will also let us know when we have added in too many predictors.

Dataset: AutoMPG

Our first dataset, the AutoMPG, is a dataset in which there are 398 cars, and based on 6 predictors (cylinders, displacement, horsepower, weight, acceleration and model year), we are attempting to predict the Miles Per Gallon that a given car gets.

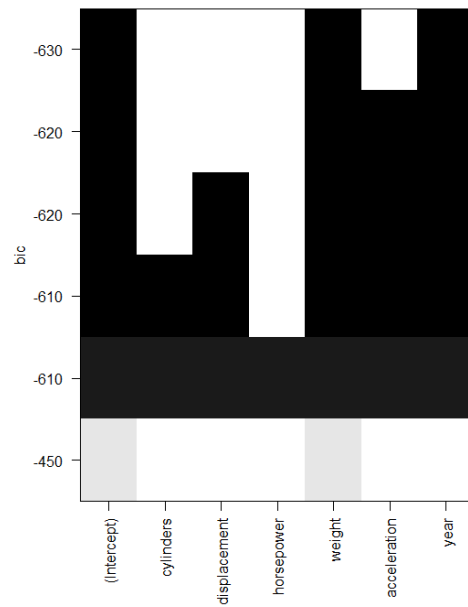
What feature selection method ended up being best:

The forward, backward, and stepwise selections performed much better than the ridge and lasso variable selection methods particularly with the MLR (with adjusted r^2 values of about 0.82) but Ridge regression ended up having the highest adjusted R-squared of all the regressions when run with a cubic regression (around 0.88)


Summary of findings:

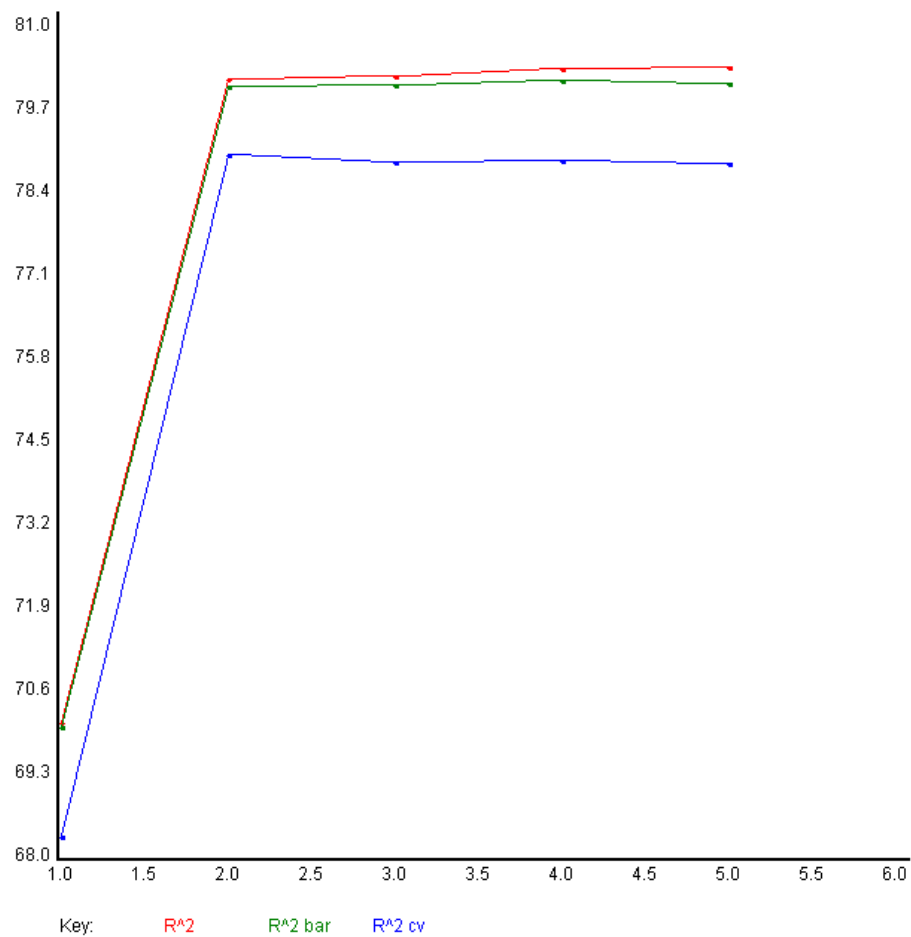
Weight and Year were selected in every selection process. All signs point to these two variables being the greatest predictors of a car's miles per gallon. This makes sense as heavier cars use more gas and cars have become more and more efficient over the years. The other predictor variables contributed in the ridge and lasso regressions, but did not contribute much to how well the model accounted for variance.

Plots:




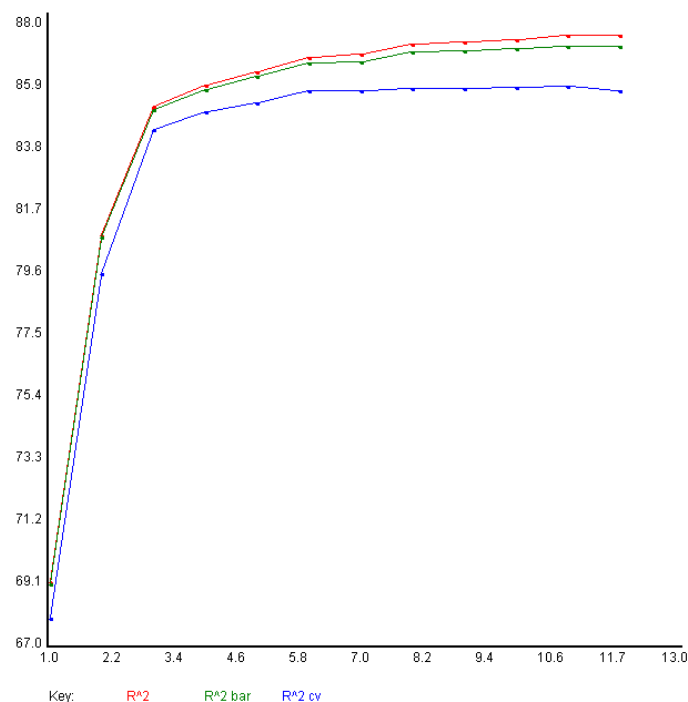
Plot of the AIC for each variable. As you can see, weight and year performed the best, as the model with only weight, year and intercept had the best bic.

 R² vs n for Regression


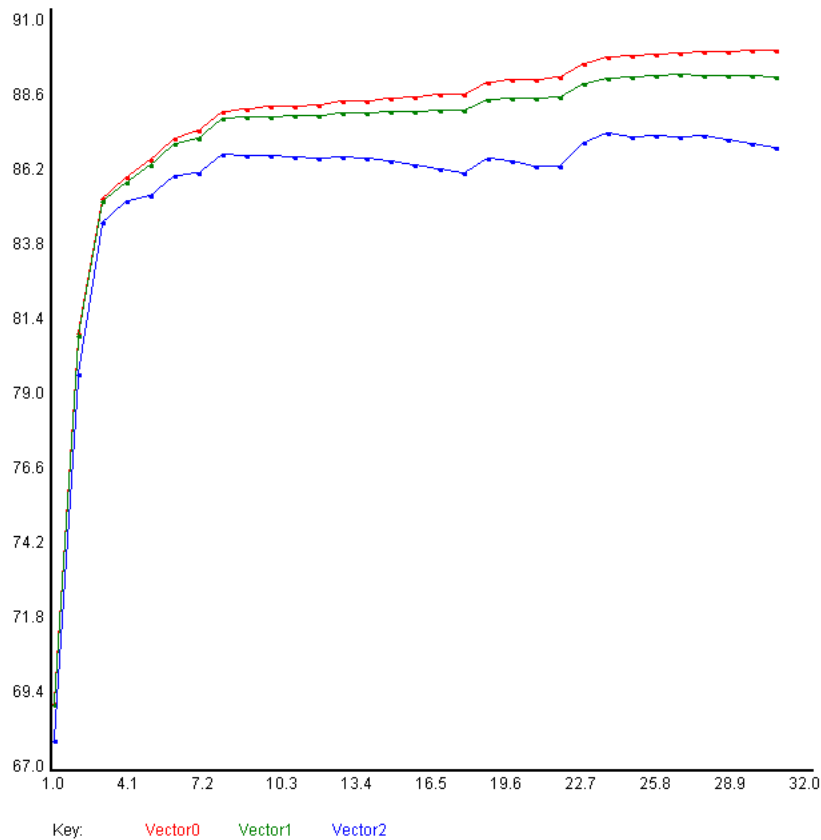


Above is the plot for the r-squared values of linear regression on the auto mpg set. As you can see, after the second variable is added, the later additions start becoming much less important.

 R² vs n for QuadRegression



Above is the plot for r-squared values for quadratic regression for the data set. As you can see, the values seem to favor a model with roughly 10 or 11 predictors, as afterwards the r² values slow down immensely in their growth.

 PlotM y_i vs. x for each i


Above is the cubic regression plot for auto mpg. This is the first plot where we truly notice a difference between the different r squared values, as r-squared cv is significantly lower than the rest. Overall though, we see very high r-squared values.

Best model:

The cubic regression using the ridge regression variables had the highest adjusted R-squared. It fit the model better than any other regression. This is no surprise as ridge regression uses each available predictor and everytime you add a variable, the R-squared will go up. Cubic regression is also the closest fit to a curved line so it accounts for more of the variance.

```
Residual standard error: 2.612 on 373 degrees of freedom
Multiple R-squared: 0.8931, Adjusted R-squared: 0.888
F-statistic: 173.2 on 18 and 373 DF, p-value: < 2.2e-16
```

Interesting observations:

The most interesting thing about this model compared to the others was the lack of effect of cross terms on the effectiveness of predictions. The models with cross terms performed about the same as the models without cross terms at higher orders, which is not typical. Otherwise, it was interesting to note that in MLR there were only 2 variables that were really important out of the 6 predictors.

Dataset: Concrete

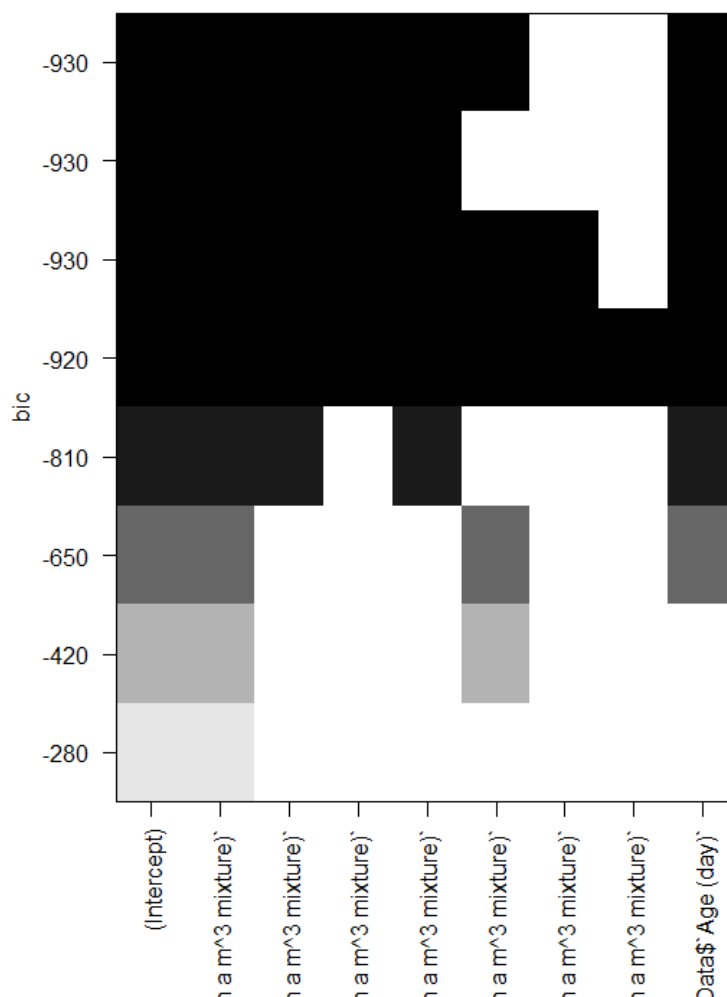
Summary of findings:

The concrete dataset has 9 predictors and 1030 instances. The dependent variable we were trying to predict was the strength of the concrete.


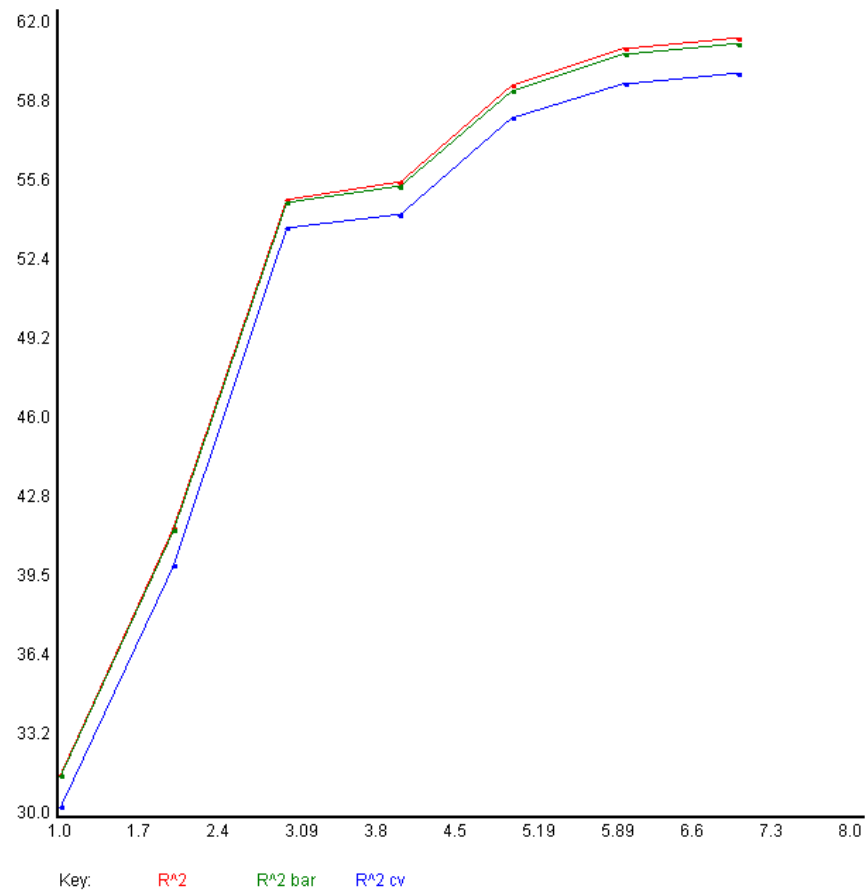
The predictors Cement, BFS, Age, Flyash, Water, and Superplasticizer were the most commonly used as each process included them in the selection.

Best Feature Selection Method:

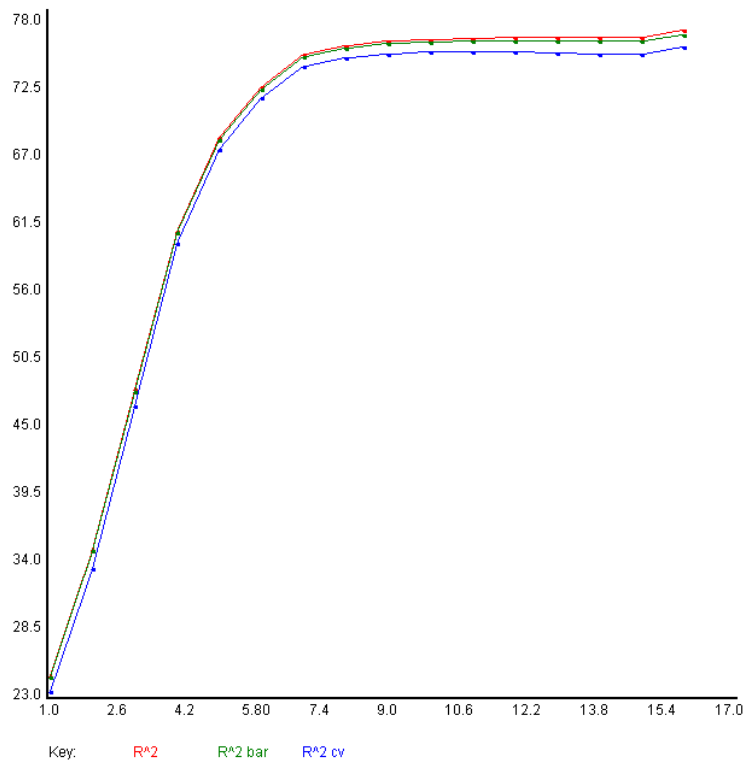
Backward Elimination and Ridge regression both had the highest adjusted R-squareds across the board. Again, this is likely because they used more variables and therefore the R-squared will go up. Overall though, all selection methods performed very well on this dataset with R-squareds in the third quadrant.

Plots:

This plot shows the aic of the model with different feature combinations. We can see that the best model is the one with all of the values except fine and coarse aggregate as predictors.

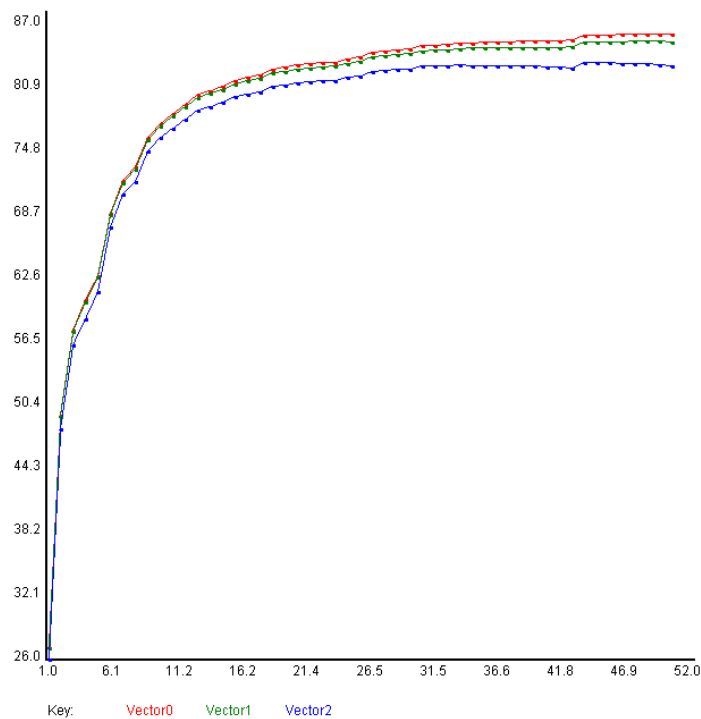
 R^2 vs n for Regression

Above is the plot for the r-squared values of linear regression on the set. As you can see, the first 5 variables added seem to be most important in predicting the concrete strength, which is consistent with the aic plot above.

R² vs n for QuadRegression

Above is the plot for r-squared values for quadratic regression for the data set. As you can see, the values seem to favor a model with roughly 7 or 8 predictors, as afterwards the r^2 values slow down in their growth.

PlotM y_i vs. x for each i



Above is the plot for r-squared values for cubic regression for the data set. The values get as high as 0.85, which is considerably greater than the values for linear regression.

Best model:

The cubic regression using the backward selection variables had the highest adjusted R-squared. It accounted for nearly 84% of the variance seen in the data.

```
Residual standard error: 6.76 on 1005 degrees of freedom  
Multiple R-squared: 0.8401, Adjusted R-squared: 0.8363  
F-statistic: 220 on 24 and 1005 DF, p-value: < 2.2e-16
```

Interesting observations:

Very good predictors in this dataset. There were high correlations across the variables so there was no one predictor that performed significantly better than the others.

Dataset: Red Wine Quality

This dataset takes in physical properties of red wines, and attempts to predict their quality ratings.

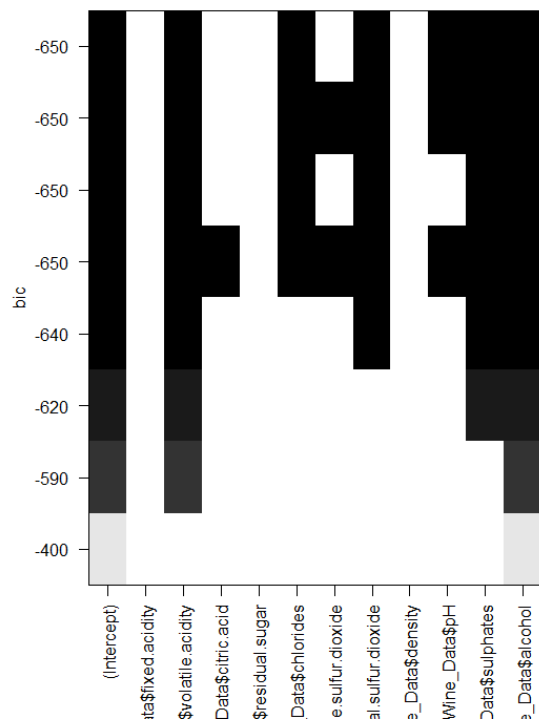
Summary of findings:

Alcohol, volatileacidity, sulphates, totalsulfur, chlorides, ph, and freesulfur were the most common predictor variables selected. However, these predictors all did a pretty poor job.

Best feature selection method:

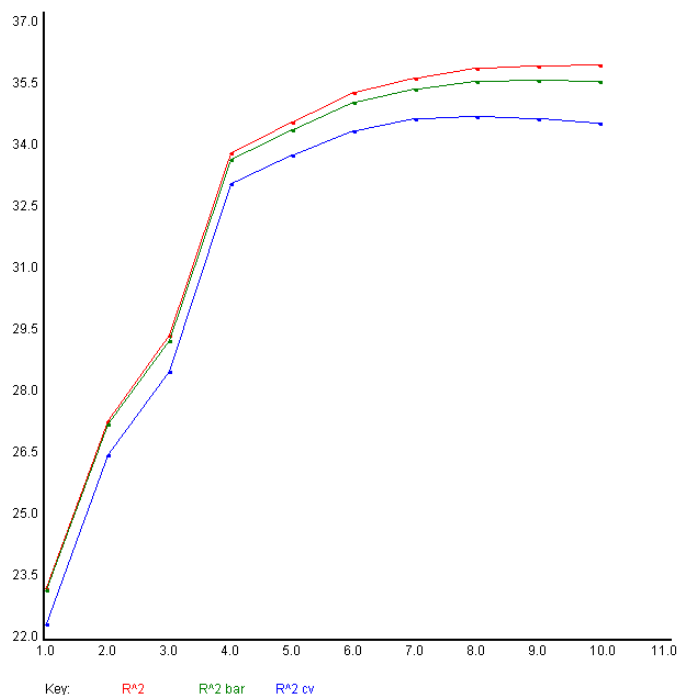
The forward, backward, and stepwise all performed the best since they used the same variables. However, all performed relatively poorly. The r-squareds were low across the board which likely means our response variable, quality, is not strongly correlated to any of our predictors.

Plots:

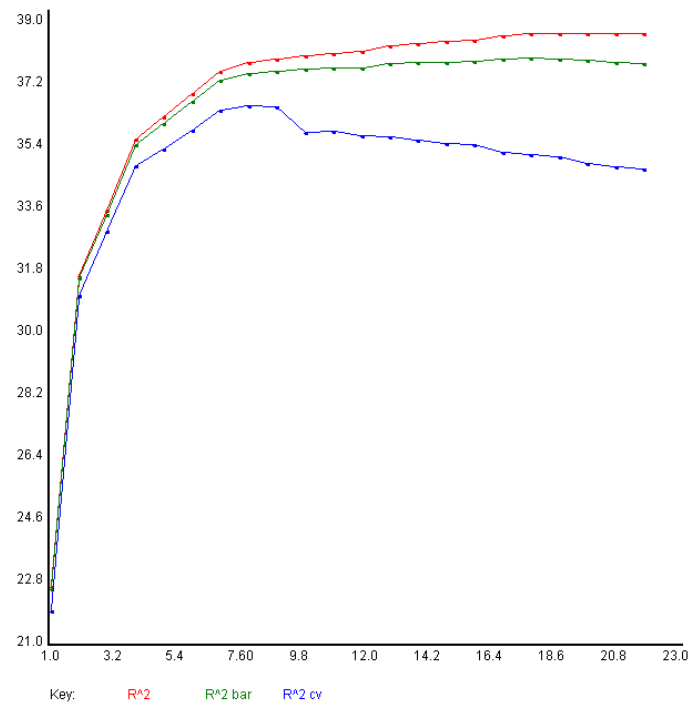


Above is the plot for bic values of the dataset. The optimal model for bic (which has a greater tendency to penalize than the r squared measures) contains only 6 predictors.

R² vs n for Regression

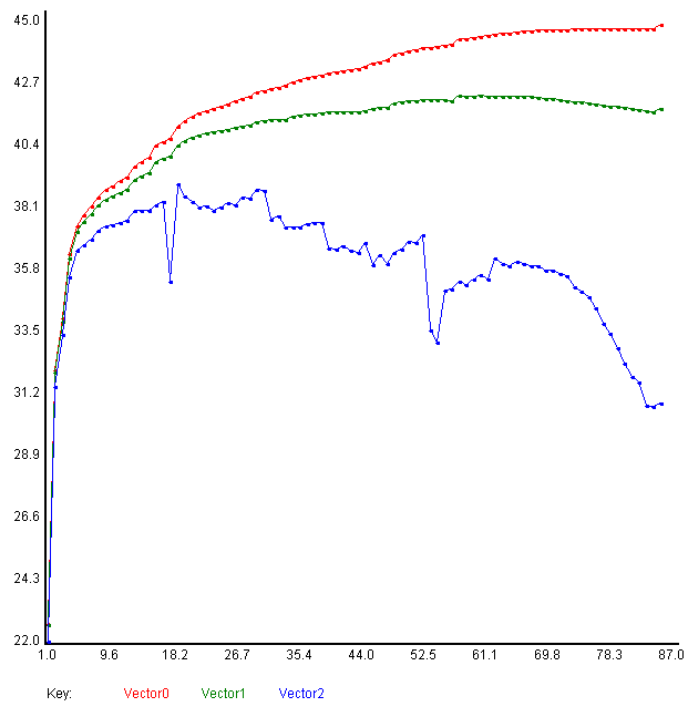


Above is the plot for r-squared values for linear regression for the data set. As you can see, the values seem to favor a model with roughly 8 predictors, as afterwards the growth slows.

R² vs n for QuadRegression

Above is the r squared graph for quadratic regression. These values are all really low and represent a less than ideal fit.

PlotM y_i vs. x for each i



Above is the r squared graph for cubic regression. Here we see a distinct difference in the different metrics, as r squared cv values are extremely low and decrease with every parameter added after 50.

Best model:

All models performed pretty poorly, but the Cubic regression with Ridge regression variables had highest adjusted R-squared, which is to be expected.

```
Residual standard error: 0.6296 on 1565 degrees of freedom
Multiple R-squared: 0.4048, Adjusted R-squared: 0.3922
F-statistic: 32.25 on 33 and 1565 DF, p-value: < 2.2e-16
```

Interesting observations:

The quality variable did not perform well as the response variable. In theory the predictors should have done a better job but that was not the case. When this happens, we can assume a lack of correlation to the response variable. Additionally, the stark contrast between r squared cv and the other metrics was of note.

Seoul Bike Rental Dataset

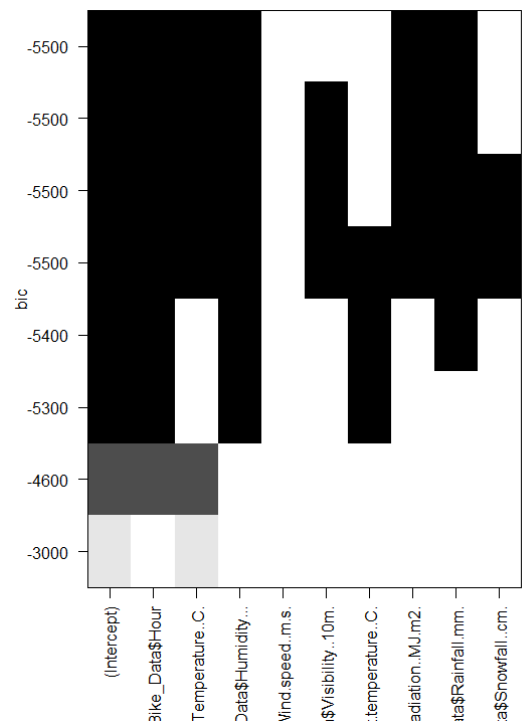
This dataset used predictors such as the weather and time to try and predict how many bikes would be rented in a given hour.

Summary of findings:


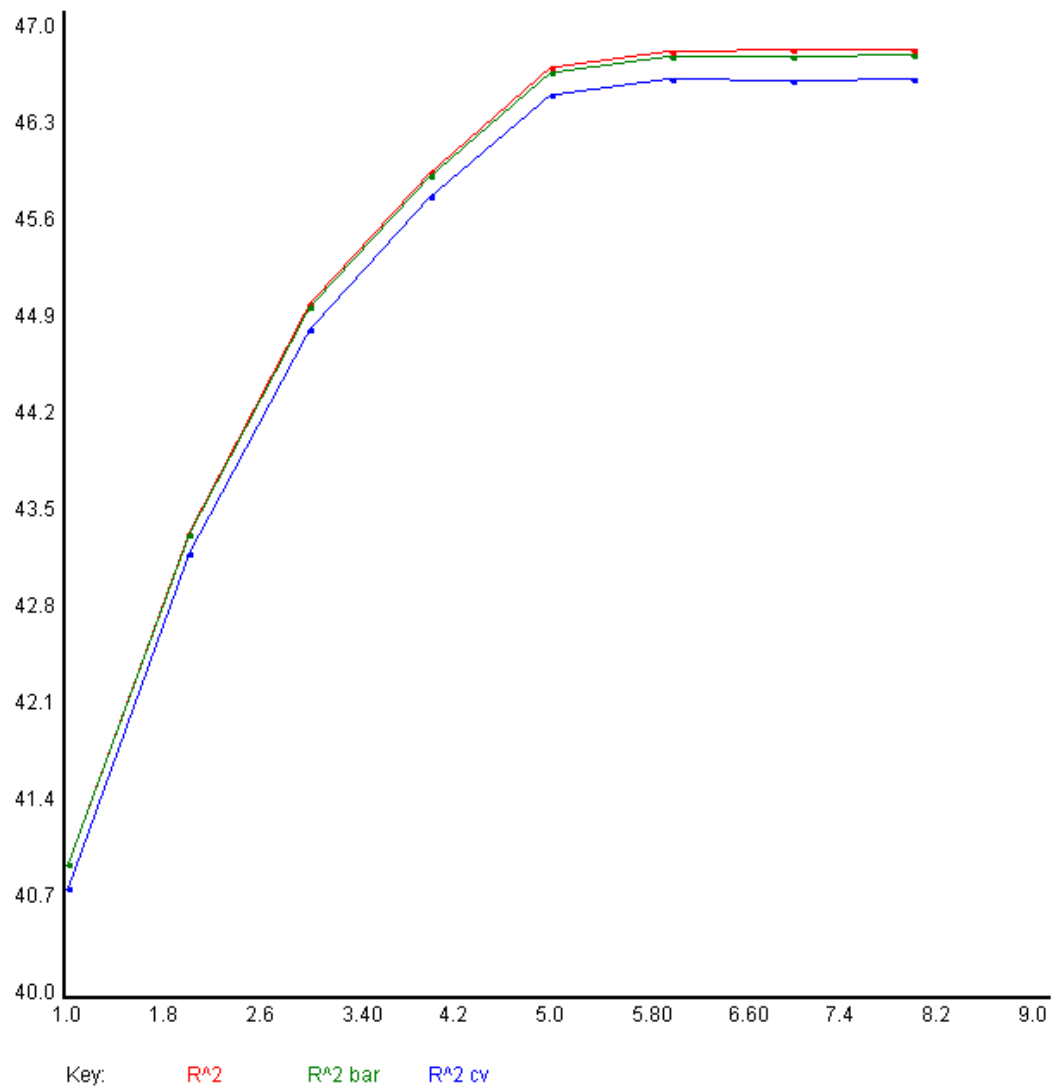
Sun and Rain had highest coefficients when running the regressions. This makes sense as these would be the biggest factors in whether or not someone rented a bike.

Best feature selection method:


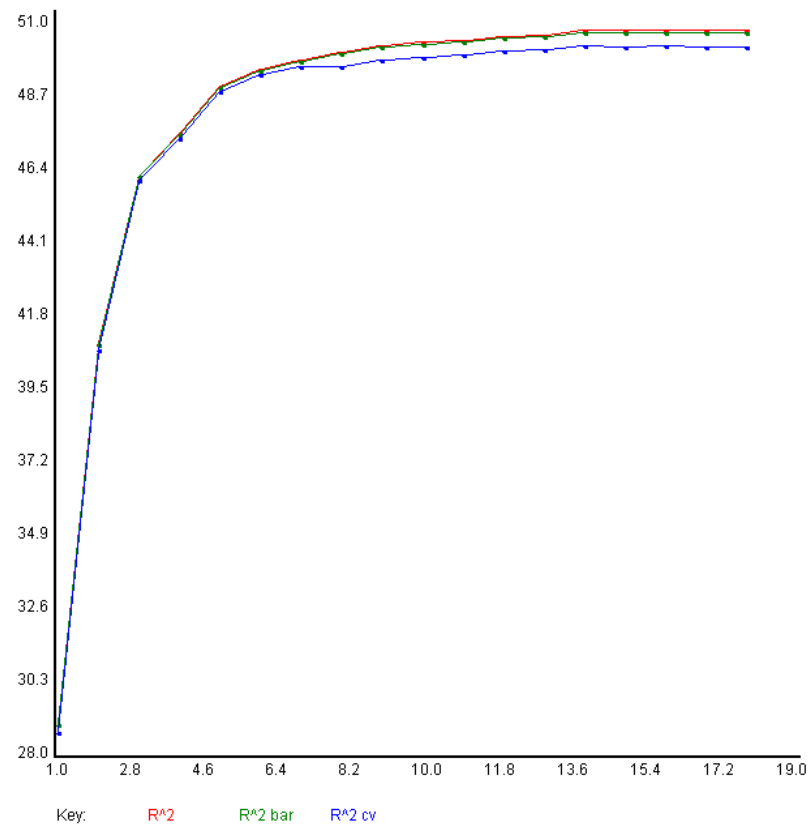
The Lasso Regression performed very well with all things considered.

Plots:

Above is a plot with bic for the dataset. This chart tells us that the best model by these standards only uses 5 predictors to fit the linear regression.

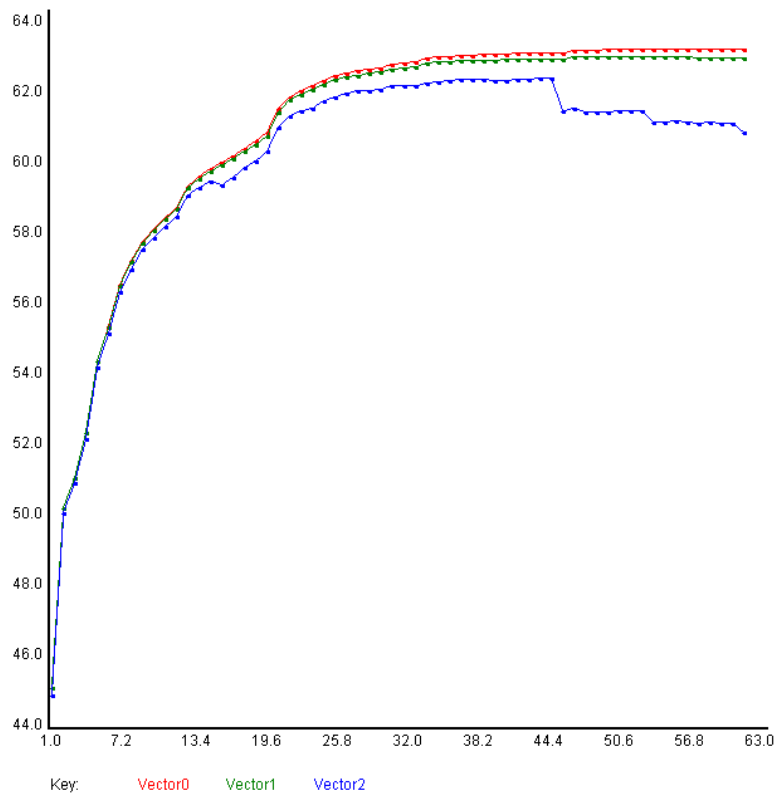
 R² vs n for Regression

Above are the r squared plots for linear regression. We can see they're about identical here, and all very low.

 R² vs n for QuadRegression

Above is the plot for r squared values of quadratic regression. Once again, they're all very similar and very low.

PlotM y_i vs. x for each i



Above is the plot for the r squared values for the dataset for cubic regression. We see a similar trend, as r squared cv dips off after about 45 terms. Otherwise, it is worth noting that these r squared values are much higher than before.

Best model:

Cubic regression using lasso variables. Once again, cubic regression fits to the data best so this is not surprising at all.

```
Residual standard error: 422.2 on 8732 degrees of freedom
Multiple R-squared: 0.5729, Adjusted R-squared: 0.5716
F-statistic: 433.8 on 27 and 8732 DF, p-value: < 2.2e-16
```

Interesting observations:

Variables did a pretty good job of predicting the response. It is a little surprising considering the real world context of the dataset.

Stability Dataset

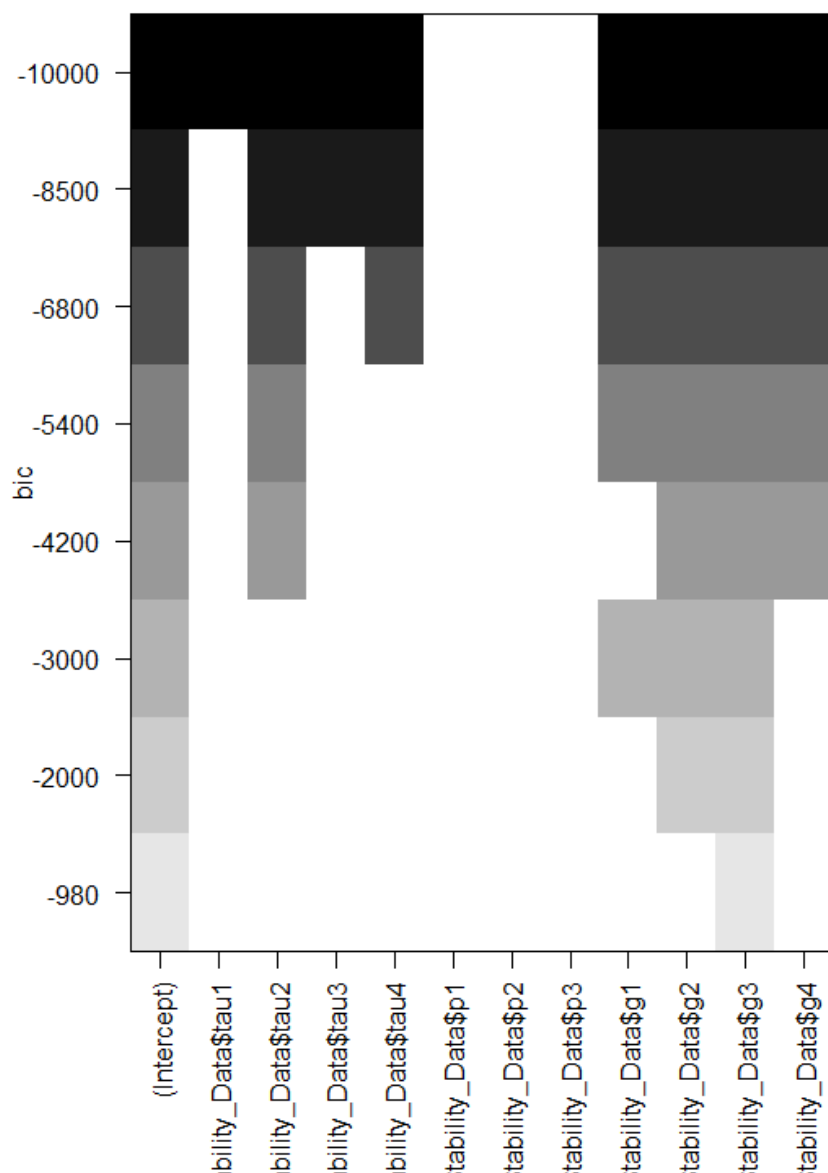
Summary of findings:

g_1 , g_2 , and g_3 seemed to be the best predictors based on the step selections. Although all variables, excluding the p variables, did a good job accounting for the variance.


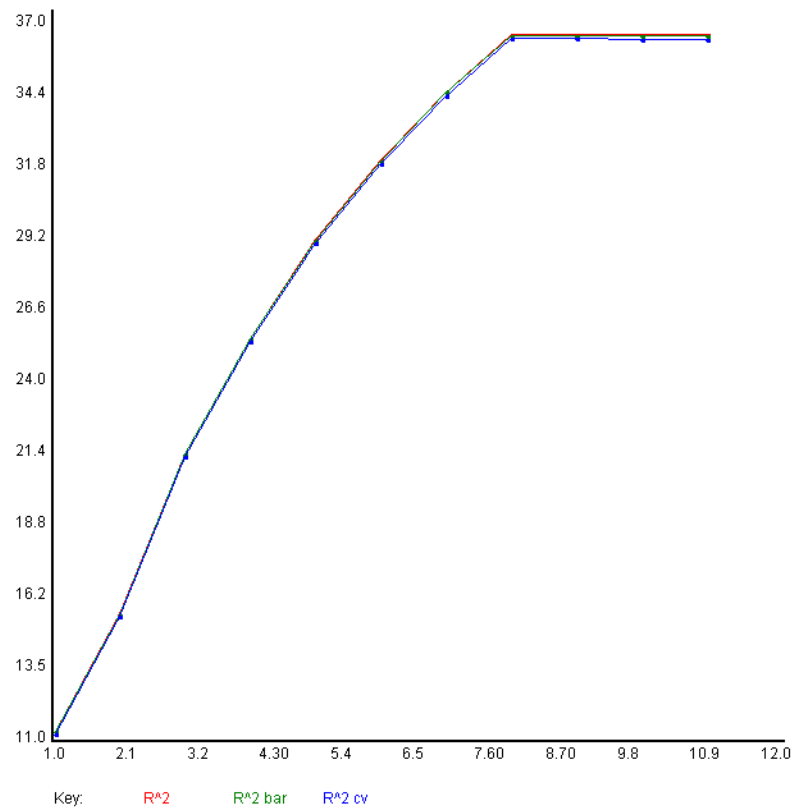
Best feature selection method:

Oddly enough, all five feature selections all performed almost the exact same with near identical R-squareds for every regression run. This is likely a testament to how similar the variable selection was, and even in the differences (including the p variables) there was almost no change.


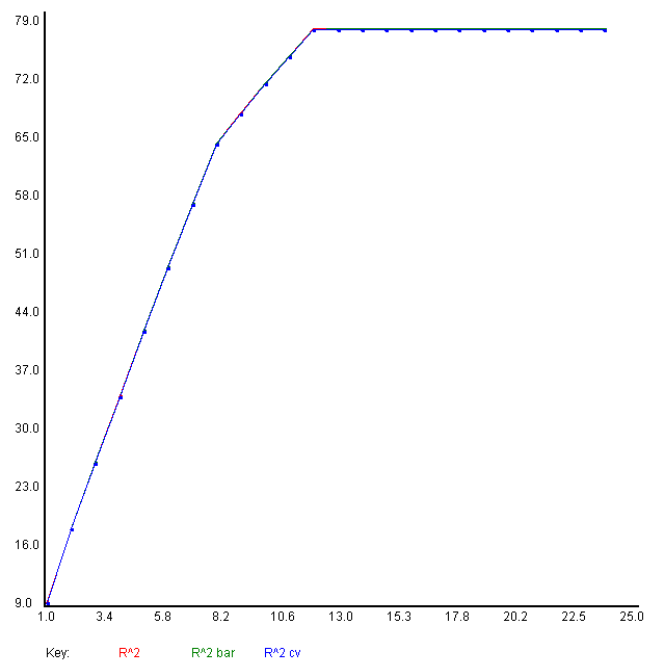
Plots:



The plot above shows us that using bic, the best model is one that contains exactly 8 of the 11 predictors.

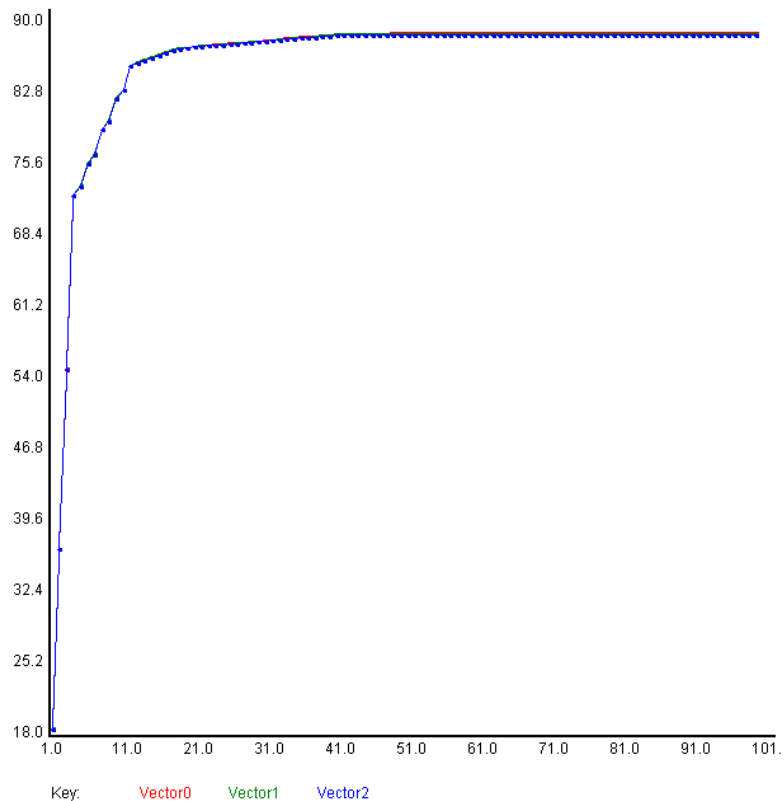
 R² vs n for Regression

Above is the plots of the r squared values for linear regression. These results are very, very poor.

 R² vs n for QuadRegression

The plot above shows quadratic regression for the dataset. The r squared values are much higher, showing that this data was much better fit to higher order terms.

PlotM y_i vs. x for each i



Above is the r squared graph for cubic regression, which hits values of close to 0.9. This shows once again that this model was much better fit to squared and cubed terms than to linear ones.

Best model:

Cubic regression using Lasso variables, but almost all the adjusted R-squareds were identical.

```
Residual standard error: 0.01716 on 9975 degrees of freedom
Multiple R-squared: 0.7845, Adjusted R-squared: 0.784
F-statistic: 1513 on 24 and 9975 DF, p-value: < 2.2e-16
```

Interesting observations:

The variable selection with this dataset had the least amount of disparity between the selections of any set we worked with.